**Description of Methods for the Interactive Workshop Measuring Language Complexity (IWMLC), Freiburg 2019**

Arturs Semenuks

*asemenuk@ucsd.edu*
*Department of Cognitive Science, UCSD, USA*

*1. Level of language addressed*

The current work is addressing the syntactic structure of language, however, in the future it is possible to extend the approach to other levels of linguistic structure as well.

*2. Measures*

I am measuring

(i) the average information density of the language at the syntactic level, more specifically the average information content of the syntactic category of each word given its preceding (syntactic) context.

(ii) the variability in the information density exhibited by the language at the syntactic level, more specifically the standard deviation of the information content of the syntactic category of each word given its preceding (syntactic) context.

*3. Operationalization and calculations of measures*

The universal part-of-speech tag (UPOS) field of the Universal Dependencies corpora is used. In each language, sentences containing symbol (SYM) or unclear part-of-speech category (X) tokens are excluded from analyses. Punctuation (PUNCT) tokens are removed from the remaining sentences.

After this, within each sentence the predictability of the syntactic category of each of its words (starting from the fourth) given their preceding syntactic context is estimated using a trigram model. First, the frequency of each of the occurring combinations

of part-of-speech categories and the preceding trigrams for part-of-speech categories are calculated in the data. Using this, the probability of the part-of-speech category occurring given its preceding trigram context is calculated. This, in turn, is used to estimate the information content of each token in the corpus using the following formula:

$$I(pos(word_i)) =$$

$$= \log\left(\frac{1}{P(pos(word_i)|pos(word_{i-1}), pos(word_{i-2}), pos(word_{i-3}))}\right)$$

where *I* is information content, $pos(word_i)$ is the part-of-speech category of a word and $word_{i-1}$, $word_{i-2}$ and $word_{i-3}$ are the preceding words.

The data on the information content of these tokens are used to calculate the average information content of the language overall, i.e. the average information density (S_idMean in Semenuks.csv), as well as the variability in the information density exhibited by language, as estimated by the standard deviation of the information content (S_idSD in Semenuks.csv).

*4. Theoretical motivation*

Are all languages equally complex? Many argue that languages do indeed differ in terms of their complexity (e.g. McWhorter, 2001; Trudgil, 2011). One of the most widely discussed hypotheses attempting to explain (some aspects of) cross-linguistic variability in complexity suggests that languages adapt to the cognitive and sociocultural niches they inhabit, i.e. languages change in complexity to accommodate the cognitive and communicative constraints of their speakers (Lupyan & Dale, 2010; Bentz & Winters, 2015).

However, many outstanding questions remain. One important question concerns the issue of psycholinguistic measures of difficulty of language acquisition, processing and production and their relation to measures of descriptive complexity – if languages

change in complexity because of their speakers' cognitive constraints, are less complex languages also indeed easier to learn, produce or understand? The question is underresearched (Miestamo, 2017): the presence or hypothesized connections between dimensions of complexity, e.g. form-meaning transparency or overspecification, and psycholinguistic difficulty still requires empirical support. Furthermore, recent experiments show that some dimensions of complexity, specifically form-to-meaning transparency, are not be always correlated with psycholinguistic difficulty (e.g. Semenuks & Berdicevskis, 2018). Thus, the question of whether simpler languages are also easier (to learn, to comprehend or to produce) requires further investigation.

One way to approach this question is to calculate complexity metrics which are more informed by psycholinguistic research and could be reasonably assumed to be transparently related to or even operationalize of some facets of learning, production or comprehension difficulty. This is what I am aiming to probe here by looking at the average information density and the variability in the information density exhibited by a language.

Information density is the amount of information transmitted per linguistic unit (e.g. word or syllable), and can be operationalized as the information content of the unit in its context, i.e. the negative log probability of the unit occurring in its context (Levy & Jaeger, 2007). For example, in the sentence "*I like coffee with milk and sugar*", the word "*like*" has a relatively high information density, as it is relatively unexpected (and so more surprising and has a higher information content) given the preceding context, whereas "*sugar*" has a relatively low information density due to its high predictability (i.e. low information content) given the rest of the sentence.

Levy and Jaeger (2007) have put forward the Uniform Information Density (UID) hypothesis, which has garnered further empirical support from subsequent research (e.g. Frank & Jaeger, 2008; Jaeger, 2010). According to the UID hypothesis, speakers aim to make the information density as uniform as possible throughout their utterances. One reason for this is that the

difficulty of processing a word has been argued to be predicted by its information density based on theoretical (e.g. Hale, 2001) and experimental (e.g. Demberg & Keller, 2008; Frank & Bod, 2011) grounds, and a UID strategy minimizes the total difficulty of processing an utterance (Levy & Jaeger, 2007).

Thus, one the one hand, it could be expected that languages overall exhibit quite similar values for the average information density and information density variability. On the other hand, it could also be expected that languages argued to be under higher pressure to be more efficiently structured, e.g. languages with more L2 speakers or speakers overall, would be under higher pressure to be more easily processed, and thus have more constrained variability of information density and higher or lower average information density values.

*5. Advantages and drawbacks of the measures*

As briefly indicated earlier, the measure is more closely tied to the psycholinguistic literature, and thus it would be easier (a) to interpret the differences between languages in this measure of complexity in terms of their effect for the speakers of the languages and (b) potentially propose transparent hypotheses for what pressures might lead to changes in thess measures, as compared to some other measures of complexity.

On the other hand, the metrics ability to be used to compare languages quite crucially depends on high quality annotated data, like the universal dependencies corpora. Furthermore, the measure relies on the possibility of a taxonomy of categories at some linguistic level that is equally applicable to all languages. For example, it isn't clear whether comparing the complexity values for different languages would be meaningful had they not all been annotated at the syntactic level using the same set of universal part-of-speech categories.

## References


Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change 3(1)*, 1–27.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *The 30th annual meeting of the Cognitive Science Society (CogSci08)* (pp. 939–944). Austin, TX: Cognitive Science Society.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science, 22*(6), 829-834.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the Second Conference of the North American chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology, 61*(1), 23-62.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (pp. 849–856). Cambridge, MA: MIT Press.

McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. *Linguistic typology, 5(2)*, 125-66.

Miestamo, M. (2017) Linguistic diversity and complexity. *Lingue e Linguaggio, 26*, 227–53.

Lupyan, G., & Dale, R. (2010) Language structure is partly determined by social structure. *PLoS ONE 5*(1), e8559.

Semenuks, A, & Berdicevskis, A. (2018) What makes a grammar difficult to learn? Experimental evidence. *EvoLang XII.*

Trudgill, P. (2011) *Sociolinguistic typology: Social determinants of linguistic complexity.* Oxford: Oxford University Press.